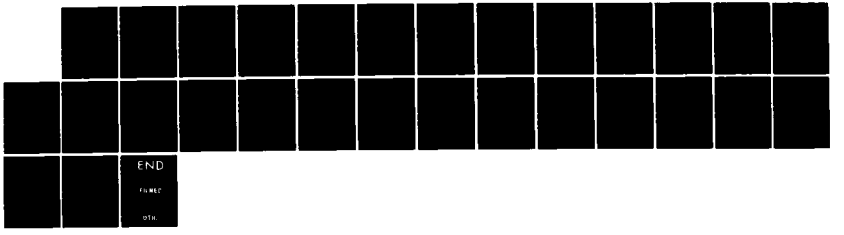
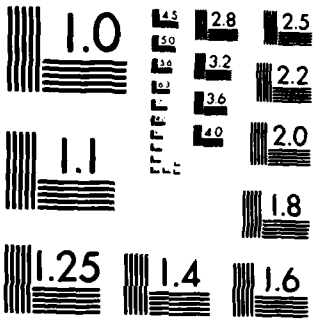
AD-A152 803 CONSIDERATIONS OF ONLINE NUMERIC DATABASES FOR SOCIAL SCIENCE RESEARCH(U) RAND CORP SANTA MONICA CA L STEWART SEP 83 RAND/P-6928 1/1

UNCLASSIFIED F/G 9/2 NL





(2)

# CONSIDERATIONS OF ONLINE NUMERIC DATABASES FOR SOCIAL SCIENCE RESEARCH

Lisa Stewart

September 1983







P-6928

## ACKNOWLEDGMENTS

I would like to thank Barbara Quint, head Reference Librarian at Rand, for so generously providing invaluable information for this paper and for making numerous corrections on the final draft.

I would also like to thank Don Trees and Jerry Koory for support in producing this paper. For their editorial services, I would like to thank Rosemary Rhoades, Rosalind Chambers, and again Don Trees.

CONTENTS

# I. INTRODUCTION

In an age of lightning-fast information, online numeric databases may seem a godsend to data users. Tantalizing with its instant information and powerful data manipulating packages, online data retrieval fulfills yesterday-is-not-too soon requests to the delight of information specialists. However, while these databases dazzle us up front, we must not forget to examine carefully basic reliability and cost issues. Items needing to be scrutinized include: data type and level availability, source identification and data equivalency, data verification, documentation, software, timeliness, and cost.

This paper is oriented to information specialists and was presented on May 21, 1983 at the IASSIST (International Association for Social Science Information Service and Technology) Conference in Philadelphia, Pennsylvania.

## II. THE ONLINE ENVIRONMENT AND THE ROLE OF THE INFORMATION SPECIALIST WITHIN IT

### DATABASE EVOLUTION

Online databases have grown substantially in the last decade. The current edition of Cuadra Associates Directory of Online Databases, an important industry source, abstracts 1,000 commercially available databases--a figure 25% higher than the previous year. Approximately half (493) of these databases are numeric in nature.(1) They have come of age, due to technological innovations and the increasing sophistication of users.

The technological innovations include:

1) increasingly inexpensive computer CPU and large amounts of quickly accessible mass storage,
2) development of large systems (e.g., IBM 3081),
3) telecommunications advancements,
4) increased real-time data collection,
5) distributed data processing and mini- and micro-computer utilization, and
6) increased availability of powerful data management systems.

In addition to these advancements, videotex and teletex will soon impact further development of numeric databases by providing access via color television sets. Just the increasingly complex nature of our society will promote the proliferation of numeric databases.

A continuing "need for quantitative data in a timely and usable form and a seemingly endless array of problem areas (e.g., energy, environment) requiring multi-disciplinary solutions and access to broad data compilations" will assure sustained growth of numeric databases.(2)

These technological innovations and the speed with which online user groups profit from them has greatly increased the size of the online market. International Resource Development says the revenues of database suppliers and distributors will exceed $1 billion in 1981 and $5.5 billion by 1991.(3) Data Resources, Inc. had about $430,000 in sales in 1980; their expected revenue for 1983 is $2.5 million. (4)

## THE INFORMATION SPECIALIST'S REACTION TO THE PROLIFERATION

One might ask why information specialists (e.g., data bank personnel and librarians) haven't gotten a hold of these databases before now if this market is booming. Although many librarians search bibliographic files online comfortably, they are frequently leery of using numeric online databases. These two online database types differ profoundly.

All around, numeric databases are more complicated. The terminology is different, the search method is different, statistical knowledge may be required, and frequently the data need manipulation. Concern about data reliability and validity often requires citations to the original source. The cost of using the numerics is frequently multiple that of bibliographic databases. These issues reflect the

real differences between source and reference databases. Whereas online bibliographic searching seems a natural extension of the information specialist's indexing skills, online numerics often require an understanding of data collection methodology and statistics.

## DATABASE DEFINITION AND TYPE

The term database means different things to different people. From an archivist's point of view, a database is a collection of data or a series of the same data collection (time-series) and its accompanying documentation--instrument, codebook, tape-layout, etc. Information specialists traditionally archive databases from a single source, collected under one methodology.

Some online databases fit this description; others do not. The specialized ones tend to follow this definition; they contain a single data collection though usually not in its entirety. Government bureaus, on the other hand, frequently produce composites, sometimes grouping several databases under one name. These large databases may group together data with similar content, geographic level, unit of analysis, or time-frame. Citibase is a good example of a numeric database of that genre. Users need to be aware of using data from multiple sources within a database because sometimes the data will not be equal. Different data collection methods will need to be verified. This sort of file is really more a data bank than a database.

By online, one means interactively accessing by telephone an external computer.

Online databases can be divided into two basic categories: reference databases and source databases. The reference databases are information intermediaries; they point to the location of the desired product. This category consists of bibliographic and referral or directory databases.

Source databases are an end-product in themselves. When people use them as such, they are sometimes referred to as "answer files." Frequently, however, they serve to produce another end-product, such as economic projections. How a user uses the data will determine what data issues are of concern to him/her.

Source databases include numeric, full-text, and textual-numeric databases. Subdividing even further, numeric databases are scientific or numeric value files and financial or socio-economic files. According to Carlos Cuadra, 85% of all source databases are of the latter category. (5) These are of the most interest to social science researchers.

Because business most aggressively pursues databases and concomitant services, commercial database services prevail in the marketplace. Government and non-profit companies do have their in-house systems, but most frequently they are not distributed externally. This paper uses commercial databases as examples.

## STRUCTURE OF THE INDUSTRY

Many of the service offerings blend into each other in this industry, making the structure hard to characterize. Basically, however, three roles divide the field: the database producer, the data distributor, and the user. Although most database producers market their databases through an online service, some sell their product directly to the user. Along with a handful of online services who produce their own databases, these firms are called integrated services.

The role of data distributor (integrated services, online, and custom information services) by itself is even more complex. While most of them provide multiple data processing services to one degree or another, the firm's main business thrust varies from company to company.

To better describe the services that the data distributors provide, we should divide the spectrum into four groups: database services, time-sharing outfits, online numeric search services, and custom information services.

1) Database services have lowest costs, fewest numeric databases, and simplest statistical capabilities. Examples are Dialog (subsidiary of Lockheed), and BRS. They emphasize products, as opposed to the expertise offered by management consulting services.

2) Time-sharing outfits traditionally made their profit on selling computer time and used to practically give away data. Now that users have their own computing capabilities with local minis and micros, these companies pay more attention to information as a revenue producer. CDC Cybernet, CDC Service Bureau, GE, and ADP Network are examples. They allow "authors" to load their databases on their dp systems, do the billing for them, and send them "royalty" checks. "Authors" are responsible for the integrity of the data, the documentation, and marketing of their databases. The time-sharing outfit serves as an intermediary, putting the user and the producer in touch with each other in case of a problem.

3) Online Numeric Search Services are top of the line. They have ornate and involved banks of data management and statistical routines, extensive documentation, full customer support, and usually very high fees. These include I. P. Sharp, Data Resources Inc., Interactive Data Corporation (Subsidiary of Citibank, produces Citibank Economic data bases containing financial data), Evan Economics (produces EEI Capsule), Rapid Data, and Compustat. The main business thrust here is expertise. Because of this some of the sophisticated online services, like DRI and Chase Econometrics, are really hybrid companies, being equally an online service and management consulting service.

Carlos Cuadra further distinguishes data distributors by the type of user they appeal to. He explains, "In both of these classes there are companies that specialize or focus on a particular market or set of users, or on particular topic areas, and other companies that take a broader approach. Lockheed provides a good example of the supermarket approach among the databases services, while General Electric exemplifies this approach among the timesharing services. These services do not focus on any one client group or on any one topic or information class, but rather draw upon the appeal of the "one-stop shopping" theme. In contrast, we see the more sharply-focused efforts of Data Resources Incorporated, which concentrates on economic information, and Interactive Data Corporation, which concentrates on financial and economic information." (6)

4) Custom Information Services arouse negative sentiment among database producers and online services. They both fear that the customizers will steal the data by copying it onto micro- and mini-computers and then compete with the other information services. Carlos Cuadra, however, feels that the customizers have been beneficial to the online market by educating users at a local level, and by being heavy users of the data themselves. He feels that end-users are worse offenders in terms of stealing data onto micros, either because they are unaware of proprietary laws or because they feel justified in doing so as customers. Cuadra comments, "Custom information

services follow scrupulous business practices if they want to stay in business." (7) The high volume of searching that customizers do gives them a high level of skill that is hard to match in-house.

## ROLE OF THE INFORMATION SPECIALIST

Information specialists, somewhat intimidated by online numerics, have been somewhat bypassed in the use of them. However, this is not inappropriate, and while they should be well-informed about these important sources of information, information specialists should not necessarily adopt a more direct role with them in the future.

End-user marketing partly explains the bypassing of the information specialist in the case of numeric databases. Strategic planners and market analysts can afford numeric databases whereas information specialists cannot. End-users also have more data understanding and computing skill than an intermediary and are thus better qualified to use the system.

Although professionally mandated to keep abreast of data availability, an active role for information specialists in utilizing numeric databases is not necessarily advisable. In many ways, numerics are beyond their capabilities. Their business is to locate data, not create it or make judgements on it. In order to obtain an answer derived from other variables, the user must perform statistical calculations. An information specialist is not and cannot be the

qualified analyst in every field, with the requisite expertise to verify the relevancy and integrity of data, massage the data into derived variables, and then analyze results before deciding upon further data treatment. To do these things, the end-user must preclude use of an intermediary and get close enough to the data to directly control it.

The real role of the information specialist is rather the traditional one--to alert end-users to information sources and their possible applications. While the information specialist may not qualify as a judge of all the issues concerning online numeric databases, it should be part of their role to provide end-users with enough information about the issues so that they may make judgements themselves.

## III. ONLINE NUMERIC DATABASE CONSIDERATIONS

The remainder of this paper centers on the databases most frequently used by social scientists (excluding economists who use macro data as frequently as micro data)--socio-economic databases.

Due to limitations of the current technology of computer mass storage, databases at this time tend be either horizontal (broad-based with shallow depth) or vertical (narrow-based with substantial depth). Many of the natural science data bases are the latter, many of the socio-economic are the former.

Because socio-economic data are both broad-based and substantially detailed, they have to be composited in order to physically fit online. Usually it is the geographic level (depth) that is sacrificed for this purpose, making them into horizontal type databases. This will present problems particular to the research it is used for.

When considering use of socio-economic data, some of the issues an information specialist may alert a user to include:

A) Type and Level of Data Available

B) Source Identification and Data Equivalency

C) Data Verification

D) Documentation

E) Software

F) Timeliness

G) Cost

(The issues above are not only limited by database type but also by data usage. Users accessing data or testing contingencies ad hoc or under strict time constraints may not be able to allow themselves the luxury of exactitude and these issues thus becomes less relevant.)

A) TYPE AND LEVEL OF DATA AVAILABLE

Obviously, a researcher will rate the utility of a database on the desirability of the type of data it provides.

Once a researcher has located the desired data, the next crucial question that will determine relevancy is usually the geographic level available. Most often researchers need data at the level lowest to collection. While many data providers recognize a user's desire to aggregate the data himself, online services determine the level of data to be loaded on the basis of commercial feasibility. Primary customers of online services need macroeconomic data for strategic planning. The most common geographic level carried online is national. Unfortunately, this is inapplicable to micro-studies which predominate in social science (excepting economic) research. Many researchers feel that lack of depth often renders the data useless to them (8).

B) SOURCE IDENTIFICATION AND DATA EQUIVALENCY

The issues of source identification and data equivalency are rather subtle. When requesting a figure online, a basic source is

referenced. It's the qualifications that are missing. Without footnotes and warnings, the data is taken out of context and its meaning shifts. Not all data from one source is produced under the same methodologies, and may not be equally reliable. Also, survey design may change from wave to wave, so even data within a time series sometimes needs to be adjusted.

Brevity offers both advantages and disadvantages to the users of online numeric systems. Users under a time constraint may not desire to deal with the details themselves. However, without complete references or a previous knowledge of the data, one may end up statistically mixing apples and oranges. Sample data is not the same as census data; survey data is not the same as administrative records. Unless users take the time to identify the source and verify the equivalency between data sources, they may invalidate the results of their statistical calculations.

Online data can sometimes assume a spurious authority due to the medium itself. "The market for numeric databases and systems is still hindered by a number of forces.... Underestimation of the importance of data evaluation is another possible obstacle to numeric database and system utilization...by being online unevaluated data may also assume an aura of authority, whereas it will be important for users to understand that this is not the case." (9)

## C) DATA VERIFICATION

Data verification concerns two issues: 1) content of the data itself and 2) methodology used to derive figures that an online service may provide for the user.

Content verification varies according to data provider. Some timesharing companies view the database producer as their client, not the end user, and leave the responsibility of data integrity to the "author." Analogous to a publisher of printed materials, timesharers feel that their duty is to correctly describe the information, not to guarantee the validity of it. Online services who offer the means to massage the data seem to assume a more active role in supporting it. I. P. Sharp, for example, employs a permanent team solely to perform range and type checks.(10) DRI reports having 25 filter programs for data verification.(11) These companies seem concerned about data validity because they realize that "one incorrect value can invalidate expensive processing time.... Collection, preparation, and validation of data for input into numeric database systems are much more expensive than for textual systems.... It is extremely important to note that online access permits individuals to use the data with little or no knowledge of their correctness, and errors resulting from bad data will discourage potential users. Such needs for error-free input will demand an increase in standards development and data formatting." (12)

To online services, the method of deriving figures is usually a trade secret. Similar to a chef in a restaurant who refuses to hand out his favorite recipes to the clientele, online services frequently refuse to divulge the methodology used to arrive at the results. Yet without knowledge of that methodology, researchers must depend solely on the reputation of the data supplier or service to backup not only the data provided but also any further analysis based on the provided information.

Many online services are sensitive to users' needs to treat the data themselves and attempt to present the data in the rawest possible form. Other companies sacrifice the pure form for a more complete data set. Example: Business International supplies macro-economic statistics profiling foreign countries' economic status. In Peter Mikelson's presentation to ONLINE '81 he reports, "Much cleaning up of data was necessary in order to provide customers with practical data. This was because the United Nations, in the course of faithfully conveying countries' reported data, provides time series with breaks, changes of base year, and other irregularities, many of which can be "massaged" without great damage to their economic significance."

He also says, "A further problem was timeliness of data. BI subscribes to hundreds of central bank and national statistical publications, many of which contain more up-to-date data than the U.N. By monitoring these additional sources and making judicious estimates when necessary, we could extend U.N. data series by a year

or two in many cases." (13) This is fine so long as they document any significant treatment of the data so that a user may verify it.

## D) DOCUMENTATION

In data documentation, as in data verification, most of the responsibility of data content falls on the author--the data producer. How much of it the online service supplies will vary.

Good documentation can resolve issues of data verification and equivalency. "Indications of good documentation include: the existence of machine-readable codebooks and data dictionaries; knowledgeable previous users; documentation and reference materials which describe sampling, data collection, processing, and analysis procedures; and the availability of ancillary materials, such as maps." (14) This documentation not only helps the user access the data, but provides a background history. The finer points such as variance and bias of sampling and non-sampling error are also important to a researcher who plans to use the data for secondary analysis.

The codebook is the most common machine-readable documentation--online or on tape. In the online environment, the codebook may or may not be available online to look up terms to search by. Provision of codebooks is usually not a problem.

Printed reports other than codebooks are much harder to come by. Data history, methodology, lists of experienced data users, or research done using the data do not accompany the data. One user felt she was very lucky if even references pointing to these items were made available.(15)

The lack of on-going historical documentation is another aspect of the documentation problem. Specifically, while online services may notify users of major changes or updates within databases for a certain time period in the online newsletter, they usually do not integrate this information permanently in the database documentation.(16)

Once again the brevity that is an advantage to some, facilitating ease of use and quick answers, proves to be a disadvantage to those social scientists seeking depth and a high degree of reliability.

E) SOFTWARE

Online services offer all different kinds of software. They range from programs which do basic statistics to highly complex forecasting models, and are complimented by no consulting to unlimited consulting. Depending on the system, a user may:

1) use a general statistical package, such as SAS or SPSS,
2) access a statistical package provided by the online service (e.g., Q-MOD by General Electric),
3) create one's own models using these packages, or
4) forecast with models provided by the online service.

In keeping with their service goals to provide quick and easy service, the interface language is user-friendly and quickly mastered. These systems are also flexible. Some meet the needs of non-data processing personnel with menu-driven programs. More sophisticated users instruct the machine with direct commands.

Although the online services don't advertise it, most of the data they market is simultaneously available either in printed form or machine-readable magnetic tapes. Software is really what online is about. "Online numeric systems work most effectively when they manipulate data. At the prices some of them charge, you may have trouble justifying them as simple data transmission systems... non-bibliographic databases contain useful information accessible at unmatched levels of retrieval pull." (17) Software makes the online services very powerful and convenient and the data worth paying more for. Instead of running several batch jobs, saving the results and merging them, one may draw across the board from several different data sources. Instead of writing your own statistical models, you can use one provided by the online services.

Needless to say, this is quite interesting to those who do not have their own computer power or those without the time or know how to do statistical manipulations. Many social scientists are not interested in external software. They tend to turn to online services for a missing piece of data, more than for software. They often already have: 1) statistical packages in-house, 2) statisticians in-house, or 3) most likely their own personal knowledge to draw upon. Some feel

that since they, not the online service, were awarded the research contract, it would be a disservice to allow someone else to do their work. Not only would they be unlikely to relinquish control of their prized analysis, but they consider online services less qualified to do it for them. They feel that the modeling they are performing is too complex, too specialized.

F) TIMELINESS

Timeliness is an online attribute that most online services take very seriously. Judith Rowe comments, "Many online statistical files contain large collections of selected time series, frequently corresponding directly to printed documents but having the advantage of incorporating weekly or monthly updates into the master file as soon as they are issued." (18)

The very sophisticated online consulting services use timeliness as an argument for charging high prices for data. As an example of timeliness, consider DRI's acquisition of the BLS Census of Wholesale Trade and the Census of Retail trade. These data are released the second Friday of every month and installed on DRI's systems by Saturday afternoon.(19)

G) COST

Cost is probably the single most influential factor in deterring use of online numeric databases for social science research. Even when data does pass all of the above qualifiers for utility, cost may prevent access.

Determining the cost of an online service (much less the cost of an individual search) is not simple. "They (online numeric databases) seem to require more negotiations, more arrangements than the traditional (bibliographic) search services.... If you don't decide to buy the tapes and load them on your own machines or to download the data from an online source and manipulate it locally (which is very popular), then you will end up using relatively simple terminals with a combination of some initial sign-up and/or annual subscription fee, plus connect-time and/or characters transmitted or CPU charge."(20) One of the most reasonably priced online services, I. P. Sharp, does not have, remarkably, start-up costs or contracts, and connect time is $1.00 per hour.(21) One of the more expensive services is Data Resources, Inc., whose average hour online costs $110 for subscribers (depending on the complexity of the session) and $155 for non-subscribers. (22)

## IV. ONLINE SERVICES CLIENT ORIENTATION

Costs clearly indicate that online industries gear their services toward the profit-making sector of business. General Electric tailors its services to businesses of the Fortune 1,000 and indicates little contact with research and academia. DRI, Inc. reports having closer ties with these two, sometimes undertaking joint ventures with the research community.

However, the online services' typical customers are million- and billion-dollar companies where big-time finances and high-stake decisions develop in an environment of multiple unknown variables, and information is of the essence. It may be worth a couple of thousand dollars to a corporation for information to make an informed decision worth millions of dollars or to acquire data quickly rather than keep high-salaried staff idle. Research and academia just are not in the same situation where time and information is of the essence. (A minor exception to this is Bid and Proposal; however, data requirements are very broad. Usually one needs only to indicate the availability of the data one would use if a grant is awarded.)

Although, currently, cost is an major deterrent to using data online, that will not be the case forever. It is only a matter of time before technological innovations resolve this and some of the other issues mentioned above. When service prices are lowered and mass

storage becomes more developed, then social scientists will become bigger online service users. The rate of change for technology coupled with the ever increasing demand for information that spurs growth in this already dynamic market suggest that this will happen not too far in the future.

## REFERENCES

(1) Fried, J. B., and G. J. Kovacs, "Numeric databases in the 80s," Drexel Library Quarterly, Summer-Fall 1982, Vol. 18, p. 3.

(2) Luedke, J., G. Kovacs, and J. Fried, "Numeric Data Bases and Systems," Annual Review of Information Science and Technology, Vol. 12, Martha E. Williams, ed., p. 119.

(3) Fried, J. B. and G. J. Kovacs, "Numeric databases in the 80s," Drexel Library Quarterly, Summer-Fall 1982, Vol. 18, p. 4.

(4,11,19,23) Interview with David D. Delany, Data Resources, Inc., April 1983.

(5,6,7,) C. A. Cuadra, Cuadra Associates, Inc., Santa Monica, CA, Proceedings, "Database Producers, Online Services, and Custom Information Services - Who Will Survive," EUSIDIC Conference Proceedings, 1978, May 1983.

(8) Interview with Kevin McCarthy and Lance Barnett, Rand Corporation researchers, May 1983.

(9,12) Luedke, J., G. Kovacs, and J. Fried, "Numeric Data Bases and Systems," Annual Review of Information Science and Technology, Vol. 12, Martha E. Williams, ed., p. 142.

(10) Interview with Andy Neilson, I. P. Sharp, April 1983.

(13) Peter Mikelson, Business International, "Design Considerations in Developing a Non-Bibliographic Database," Online '81 Conference Proceedings, Dallas, Texas, November 1981.

(14) McGee, J. and D. P. Trees, "Major Available Social Science Machine-Readable Databases," Drexel Library Quarterly, Summer-Fall 1982, Vol. 18, p. 114.

(15,16,17,20) Barbara Quint, Rand Corporation, "The Librarian Versus the Non-Bibliographic Online Database: The Next Round," Annual Conference of the International Association for Marine Science Libraries and Information Centers, November 1981.

(18) Judith Rowe, Princeton University Center, "Machine-Readable Data Files of Government Publications," Government Publications Review, 1981, Vol. 8A, p. 343.

(21) I. P. Sharp Associates, Inc., Sharp APL System Usage Agreement.

(22) Data Resources, Inc., conversation with John Cala, September 1983.

## BIBLIOGRAPHY

Berger, M. C., "An Update: Online Non-Bibliographic Databases," Cuadra Associates, Online '81, Dallas, Texas.

Directory of Data Firms, American Demographics, February 1983, Vol. 5 No. 2, pp. 30-39.

Directory of On-Line Databases, Vol. 4, No. 1, Cuadra Associates, Santa Monica.

Cuadra, C. A., "Database Producers, Online Services, and Custom Information Services - Who Will Survive?," Cuadra Associates, Inc., Santa Monica, California

Hawkins, D. T., "Online Information Retrieval Bibliography - Fifth Update," Online Review, 1982, Vol. 6, No. 2, pp. 147-208.

Mikelson, P., "Design Considerations in Developing a Non-Bibliographic Database," Online '81, Dallas Texas, Business International.

"Numeric Databases," Drexel Library Quarterly, Summer-Fall 1982, Drexel University, Philadelphia.

Rowe, J., "On-Line Sources of Government-Produced Data, Part II - Numeric Data," Government Publications Review, Vol. 8A, pp. 343-351, 1981.

END

FILMED

5-85

DTIC